Integrating Face and Gait for Human Recognition

Xiaoli Zhou and Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside
Riverside CA 92521
{xzhou, bhanu}@vislab.ucr.edu

Abstract

This paper introduces a new video based recognition method to recognize non-cooperating individuals at a distance in video, who expose side views to the camera. Information from two biometric sources, side face and gait, is utilized and integrated for recognition. For side face, we construct Enhanced Side Face Image (ESFI), a higher resolution image compared with the image directly obtained from a single video frame, to fuse information of face from multiple video frames. For gait, we use Gait Energy Image (GEI), a spatio-temporal compact representation of gait in video, to characterize human walking properties. The features of face and the features of gait are obtained separately using Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA) combined method from ESFI and GEI, respectively. They are then integrated at match score level. Our approach is tested on a database of video sequences corresponding to 46 people. The different fusion methods are compared and analyzed. The experimental results show that (a) Integrated information from side face and gait is effective for human recognition in video; (b) The idea of constructing ESFI from multiple frames is promising for human recognition in video and better face features are extracted from ESFI compared to those from original face images.

1. Introduction

It has been found to be very difficult to recognize a person from arbitrary views when one is walking at a distance. For optimal performance, a system should use as much information as possible from the observations. A fusion system, which combines face and gait cues from video sequences, is a potential approach to accomplish the task of human recognition.

The most general solution to analyze face and gait video data from arbitrary views is to estimate 3-D models. However, the problem of building reliable 3-D models for non-rigid face with flexible neck and the articulated human body from low resolution video data remains a hard one. In recent years, integrated face and gait recognition approaches without resorting to 3-D models have achieved some progress [10] [13] [14] [16].

Most current gait recognition algorithms rely on the availability of the side view of the subject since human gait or the style of walking is best exposed when one presents a side view to the camera. For face recognition, on the other hand, it is preferred to have frontal views. These conflicting requirements pose some challenges when one attempts to integrate face and gait biometrics in real world applications. In the previous fusion systems [10] [13] [14], the side view of gait and the frontal view of face are used. In [10], Kale et al. present a gait recognition algorithm and a face recognition algorithm based on sequential importance sampling. The database contains video sequences for 30 subjects walking in a single camera scenario. For face recognition, only the final segment of the database can present a nearly frontal view of face and it is used as the probe. The galley consists of static faces for the 30 subjects. Therefore, they have to perform still-to-video face recognition. In [13] [14], Shakhnarovich et al. compute an image-based visual hull from a set of monocular views of multiple cameras. It is then used to render virtual canonical views for tracking and recognition. They discuss the issues of cross-modal correlation and score transformations for different modalities and present the cross-modal fusion. In their work, 4 monocular cameras must be used to get both the side view of gait and the frontal view of face simultaneously. Recently, Zhou et al. propose a system [16], which combines cues of face profile and gait silhouette from the single camera video sequences. It is based on the fact that a side view of face is more likely to be seen than a frontal view of face when one exposes the best side view of gait to the camera. The data is collected for 46 people with 2 video sequences per person. Even though face profile in Zhou et al.'s work is used reasonably, it only contains shape

Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)
0-7695-2646-2/06 $20.00 © 2006 IEEE

Table 1. Our approach for integrating face and gait for human recognition vs. the previous work

Features	Kale et al. [10]	Shakhnarovich et al. [13] [14]	Zhou et al. [16]	This Paper
Biometrics	• Frontal face • Gait	• Frontal face • Gait	• Face profile • Gait	• Side face • Gait
Number of Cameras	1	4	1	1
Face Features and Recognition	• Motion vectors • Time series model • Posterior distribution • MAP	• PCA features of the detected face. • k-NN	• Curvature based features of face profile from the high-resolution image. • Dynamic time warping	• Face features of Enhanced Side Face Image (ESFI) • PCA and MDA combined method • k-NN
Gait Features and Recognition	• Entire canonical view image • Template matching based on dynamic time warping.	• Means and standard deviation of moments, and centroid. • k-NN	• Entire Gait Energy Image (GEI) • Template matching	• Gait features of Gait Energy Image (GEI) • PCA and MDA combined method • k-NN
Data	• 30 subjects • Number of sequences per person: not specified. • Static images as the face gallery	• 26 subjects [13] • 2 to 14 sequences per person [13] • 12 subjects [14] • 2 to 6 sequences per person [14]	• 14 subjects • 2 sequences per person	• 46 subjects • 2 sequences per person
Fusion Methods	• Hierarchical fusion • Sum/Product rule	• Min, Max, Sum and Product rules [13]. • Sum rule [14]	• Hierarchical fusion • Sum and Product rules	• Max, Sum and Product rules

information of the side view of face and misses its intensity information. In this paper, an innovative video based fusion system is proposed, aiming at recognizing non-cooperating individuals at a distance in a single camera scenario. Information from two biometric sources, side face and gait, from the single camera video sequence, is combined. Side face, not face profile, includes entire side views of eye, nose and mouth, possessing both shape information and intensity information. Therefore, it has more discriminating power for recognition.

Table 1 presents a summary of related work and compares it with the work presented in this paper. It is difficult to get reliable information of a side face directly from a video frame for recognition task because of limited resolution. To overcome this problem, we construct Enhanced Side Face Image (ESFI), a higher resolution image compared with the image directly obtained from a single video frame, to fuse information of face from multiple video frames. The idea relies on the fact that the temporally adjacent frames in a video sequence, in which one is walking with a side view to the camera, contain slightly different, but unique information of a side face [16]. Experiments show that better face features can be extracted from constructed ESFI compared to those from original side face images.

In this paper, face and gait information are fused at match score level using different fusion schemes. Face features and gait features are obtained separately using PCA and MDA combined method from ESFI and Gait Energy Images (GEI), respectively. Experiments are implemented to compare the performance between different biometrics and different fusion methods. Performance analyses are presented in detail.

The paper is organized as follows. Section 2 introduces a video based fusion system, utilizing and integrating information from side face and gait. It explains the construction of ESFI and GEI, and feature representation using PCA and MDA combined method. It also presents an approach to generate synthetic match scores for fusion and a description of the classification methods. In Section 3, a number of dynamic video sequences are tested using the approach presented. Experimental results are compared and discussed. Section 4 concludes the paper.

2. Technical Approach

The overall technical approach is shown in Figure 1. We first construct Enhanced Side Face Image (ESFI) as the face template and Gait Energy Image (GEI) as the gait template from video sequences. In the training procedure, we perform a component and discriminant analysis separately on face templates and gait templates obtained from all training videos. As a result, transformation matrices and features that form feature gallery are obtained. In the recognition procedure, each testing video is processed to generate both face templates and gait templates, which are transformed by transformation matrices to obtain face features and gait features, respectively. These testing features are compared with gallery features in the database, and then different fusion strategies are applied to combine the results of face classifier and the results of gait classifier to improve recognition performance.

2.1. Enhanced Side Face Image Construction

Multiframe resolution enhancement seeks to construct a single high-resolution image from several low-resolution images. These low-resolution images must be of the same object, taken from slightly different angles, but not so much as to change the overall appearance of the object in the im-

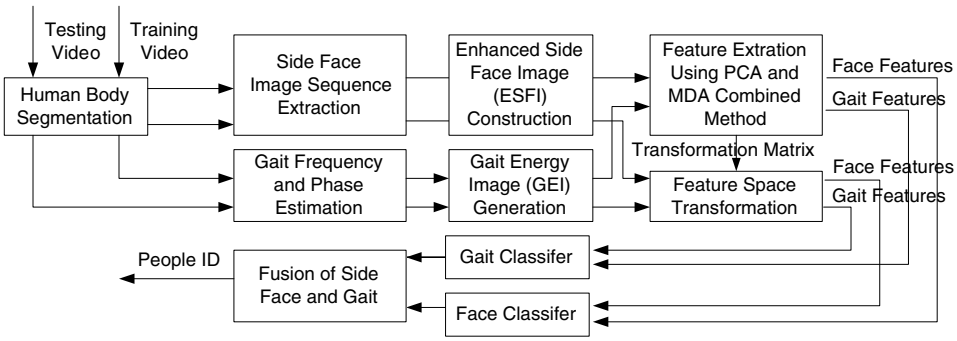


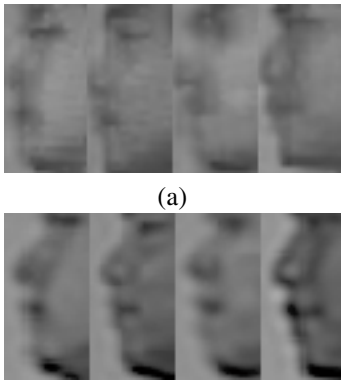
Figure 1. Technical approach for integrating side face and gait in video.

age.

We use a simple background subtraction method [5] for human body segmentation. A human body can be divided into two parts according to the proportion of its parts [7]: from the top of the head to the bottom of the chin, and then from the bottom of the chin to the bottom of the foot. Human head is defined as the part from the top of the head to the bottom of the chin. Considering the height of hair and the length of neck, we assume that the upper 16% of the segmented human body includes the human head. In this paper, original low-resolution side face images are first localized and extracted by cutting the upper 16% of the segmented human body obtained from multiple video frames. Then an iterative method [9] is applied to construct a high-resolution side face image from the aligned low-resolution side face images. Figure 2 shows one low-resolution face image and one reconstructed high-resolution face image. For comparison, we resize the low-resolution face image using bilinear interpolation.

Figure 2. Resized low-resolution face image (left) and constructed high-resolution face image (right).

Before feature extraction, all high-resolution side face images are normalized. The normalization is based on the locations of nasion, pronasale and throat on the face profile. These three fiducial points are identified by using a curvature based fiducial extraction method [2]. In this paper, Enhanced Side Face Image (ESFI) is a subimage, excluding hair and background, obtained from normalized versions of the high-resolution side face images. Similarly, Original Side Face Image (OSFI) is a subimage, excluding hair and background, obtained from normalized versions of the low-resolution side face images. Examples of resized OSFIs and ESFIs for 4 people are shown for comparison in Figure 3. Clearly, ESFIs have better quality than OSFIs.



(a)

(b)

Figure 3. Examples of 4 people: (a) Resized OSFIs (b) ESFIs

2.2. Gait Energy Image Construction

Regular human walking can be considered as cyclic motion where human motion repeats at a stable frequency [6]. Therefore, it is possible to divide the entire gait sequence into cycles. Since human body segmentation is performed on original human walking sequences, we begin with the extracted binary silhouette image sequences. The silhouette preprocessing includes size normalization and horizontal alignment. In a preprocessed silhouette sequence, the time series signal of lower half silhouette size from each frame indicates the gait frequency and phase information. We estimate the gait frequency and phase by maximum entropy spectrum estimation [12] from the time series signal.

Given the preprocessed binary gait silhouette image $B_t(x, y)$ at time t in a sequence, the grey-level gait energy image (GEI) is defined as follows [6]:

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B_t(x, y) \quad (1)$$

where N is the number of frames in the complete cycle(s) of a silhouette sequence, t is the frame number of the sequence (moment of time), and x and y are values in the

Figure 4. Examples of normalized and aligned silhouette images in a gait cycle. The right most image is the corresponding gait energy image (GEI).

2D image coordinate. Figure 4 shows the sample silhouette images in a gait cycle and the right most image is the corresponding GEI. As expected, GEI reflects major shapes of silhouettes and their changes over the gait cycle. GEI has several advantages over the gait representation of binary silhouette sequence. GEI is not sensitive to incidental silhouette errors in individual frames. Moreover, with such a 2D template, we do not need to consider the time moment of each frame, and the incurred errors can be, therefore, avoided.

2.3. Feature Extraction Using PCA and MDA Combined Method

Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA) have long been used for the appearance-based object recognition [1][8]. PCA reduces the dimension of feature space, and MDA automatically identifies the most discriminating features. In this paper, PCA and MDA combined method is applied to face templates ESFIs and gait templates GEIs separately to get low dimensional feature representation for side face and gait.

Let $X \in R^N$ be a random vector representing an ESFI or a GEI, where N is the dimensionality of the corresponding image. The covariance matrix of X is defined as $\Sigma_x = E([x - E(x)][x - E(x)]^T)$, where $E(\cdot)$ is the expectation operator and T denotes the transpose operation. The covariance matrix Σ_x can be factorized into the following form:

$$\Sigma_x = \Phi \Lambda \Phi \tag{2}$$

where $\Phi = [\Phi_1 \Phi_2 ... \Phi_N] \in R^{N \times N}$ is the orthogonal eigenvector matrix of Σ_x ; $\Lambda = \{\lambda_1 \lambda_2 ... \lambda_N\} \in R^{N \times N}$ is the diagonal eigenvalue matrix of $\Sigma_x$with diagonal elements in descending order ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$). One important property of PCA is its optimal signal reconstruction in the sense of minimum mean square error (MSE) when only a subset of principal components are used to represent the original signal. An immediate application of this property is the dimensionality reduction:

$$Y = P_{pca}^T X \tag{3}$$

where $P_{pca} = [\Phi_1 \Phi_2 ... \Phi_m], m < N$. The lower dimensional vector $Y \in R^m$ captures the most expressive features of the original data X.

MDA seeks a transformation matrix W that maximizes the ratio of the between-class scatter matrix S_B to the within-class scatter matrix S_W : $J(W) = \frac{|W^T S_B W|}{|W^T S_W W|}$. Suppose that $w_1, w_2, ...w_c$ and $n_1, n_2, ...n_c$ denote the classes and the number of images within each class, respectively, with $n = n_1 + n_2 + ... + n_c$ and $w = w_1 \cup w_2 \cup ... \cup w_c$. c is the number of classes. The within-class scatter matrix is $S_W = \sum_{i=1}^{c} \sum_{Y \in w_i} (Y - M_i)(Y - M_i)^T$ and the between-class scatter matrix is $S_B = \sum_{i=1}^{c} n_i (M_i - M)(M_i - M)^T$. where $M_i = \frac{1}{n_i} \sum_{Y \in w_i} Y$ and $M = \frac{1}{n} \sum_{Y \in w} Y$ are the means of the class i and the grand mean. $J(W)$ is maximized when the columns of W are the generalized eigenvectors of S_B and S_W corresponding to the largest generalized eigenvalues in

$$S_B \Psi_i = \lambda_i S_W \Psi_i \tag{4}$$

There are no more than $c - 1$ nonzero eigenvalues λ_i and the corresponding eigenvectors Ψ_i. The transformed feature vector is obtained as follows:

$$Z = P_{mda}^T Y = P_{mda}^T P_{pca}^T X = QX \tag{5}$$

where $P_{mda} = [\Psi_1 \Psi_2 ... \Psi_k], k < c$ and Q is the overall transformation matrix. We can choose k to perform feature selection and dimensionality reduction. The choice of the range of PCA and the dimension of MDA reflects both the energy need and the magnitude requirement [4]. The lower dimensional vector $Z \in R^k$ captures the most expressive and discriminating features of the original data X.

2.4. Recognition by Integrating ESFI and GEI

Given a testing video, we obtain low dimensional feature vectors $F' = Q^f F$ and $G' = Q^g G$ from the face template F and the gait template G, respectively, by using PCA and MDA combined method as Equation (5). Q^f and Q^g are the overall transformation matrices for face and gait, respectively. The Euclidean distance for the face classifier and the gait classifier are obtained as

$$\begin{aligned} D_i^F &= ||F' - U_i^F|| \\ D_i^G &= ||G' - U_i^G|| \end{aligned} \tag{6}$$

where U_i^F and U_i^G, $i = 1, 2, ...c$, are the prototypes of class i for face and gait, respectively. Before fusion, it is necessary to map distances obtained from the different classi-

Figure 5. Examples of video sequences.

fiers to the same range of values. We use exponential transformation here. Given that the distance for a probe X are $S_1, S_2...S_c$, we obtain the normalized match scores

$$\hat{S}_i = \frac{exp(-S_i)}{\sum_{i=1}^{c} exp(-S_i)} \quad i = 1, 2, ...c. \tag{7}$$

After normalization, the match scores of face templates and the match scores of gait templates from the same class are fused based on different fusion methods. Since face and gait can be regraded as two independent biometrics in our scenario, synchronization is totally unnecessary for them. To take advantage of information for a walking person in video, we use all the possible combinations of face match scores and gait match scores to generate new match scores, which encode information from both face and gait. The new match scores are called *synthetic match scores*. It is reasonable to generate synthetic match scores in this way, since ESFI is built from multiple video frames and gait energy image (GEI) is a compact spatio-temporal representation of gait in video. In this paper, we use 2 face match scores and 2 gait match scores to generate 4 synthetic match scores for one person from each video.

Distances representing dissimilarity become match scores representing similarity by using Equation (7), so the unknown person should be classified to the class for which the synthetic match score is the largest. Let $\hat{D}_i^F$ and $\hat{U}_i^G$ be the normalized match scores of D_i^F and U_i^G, respectively. The unknown person is classified to class k if

$$R\{\hat{D}_k^F, \hat{D}_k^G\} = \max R\{\hat{D}_i^F, \hat{D}_i^G\} \tag{8}$$

where $R\{,\}$ means a fusion method. In this paper, we use SUM, PRODUCT and MAX rules. Since we obtain more than one synthetic match scores after fusion for one testing video sequence. Equation (8) means the unknown person is classified to the class which gets the maximum synthetic match score out of all the synthetic match scores corresponding to all the classes.

3. Experimental Results

3.1. Experiments

The data is obtained by Sony DCR-VX1000 digital video camera recorder. We collect 92 video sequences of 46 people walking outside and exposing a side view to the camera, at 30 frames per second. The resolution of each frame is 720x480. The distance between people and the video camera is about 10 feet. Each person has two video sequences, one for training and the other one for testing. Each video sequence includes one person. Figure 5 shows some examples of the data.

For gait, we obtain 2 complete walking cycles from a video sequence according to the gait frequency and gait phase. Each walking cycle includes about 20 frames. We construct 2 GEIs corresponding to 2 walking cycles from one video sequence. The resolution of each GEI is 300x200. For face, we also construct 2 high-resolution side face images from one video sequence. Each high-resolution side face image is built from 10 low-resolution side face images that are extracted from 10 adjacent video frames. The resolution of low-resolution side face images is 70x70 and the resolution of reconstructed high-resolution side face images is 140x140. After normalization as in Section 2.1, the resolution of ESFI is 64x32. For 46 people, we obtain 92 ESFIs and 92 GEIs as the gallery and another 92 ESFIs and 92 GEIs as the probe. After feature extraction using PCA and MDA combined method as described in Section 2.3, we have 92 face feature vectors and 92 gait feature vectors in the gallery and another 92 face feature vectors and 92 gait feature vectors in the probe. The dimensionality of face features is 25 and the dimensionality of gait features is 15. For fusion, as explained in Section 2.4, we generate 4 synthetic match scores based on 2 face match scores and 2 gait match scores for one person from each video. Totally, we have 184 synthetic match scores corresponding to 46 people in the gallery and 184 synthetic match scores corresponding to 46 people in the probe.

Recognition performance is used to evaluate our method. It is defined as the ratio of the number of the correctly recognized people to the number of all the people. Table 2 shows the performance of individual biometric. Table 3 shows the performance of fusion using different combination rules. We name 46 people from No. 1 to No. 46. In Table 2 and Table 3, the error index gives the number of misclassified sequence. For comparison, we also show the performance using face features from Original Side Face Image (OSFI) to demonstrate the performance improvement by using constructed ESFI. The resolution of OSFI is 34x18. The procedures of feature extraction, synthetic match score genera-

Table 2. Performance and error index of individual biometric

Fusion Method	Original Face (OSFI)	Enhanced Face (ESFI)	Gait (GEI)
Recognition Rate	71.7%	84.8%	87.0%
Error Index	2 4 6 8 13 18 20 22 26 28 35 44 46	2 4 8 13 16 36 46	2 5 6 8 9 13

Table 3. Performance and error index of fusion

Fusion Method		Sum Rule	Product Rule	Max Rule
OSFI & GEI	Recognition Rate	89.1%	84.8%	84.8%
	Error Index	2 6 8 9 13	2 4 5 6 8 9 13	2 6 8 9 13 26 46
ESFI & GEI	Recognition Rate	91.3%	84.8%	91.3%
	Error Index	2 8 9 13	2 4 5 6 8 9 13	2 8 13 46

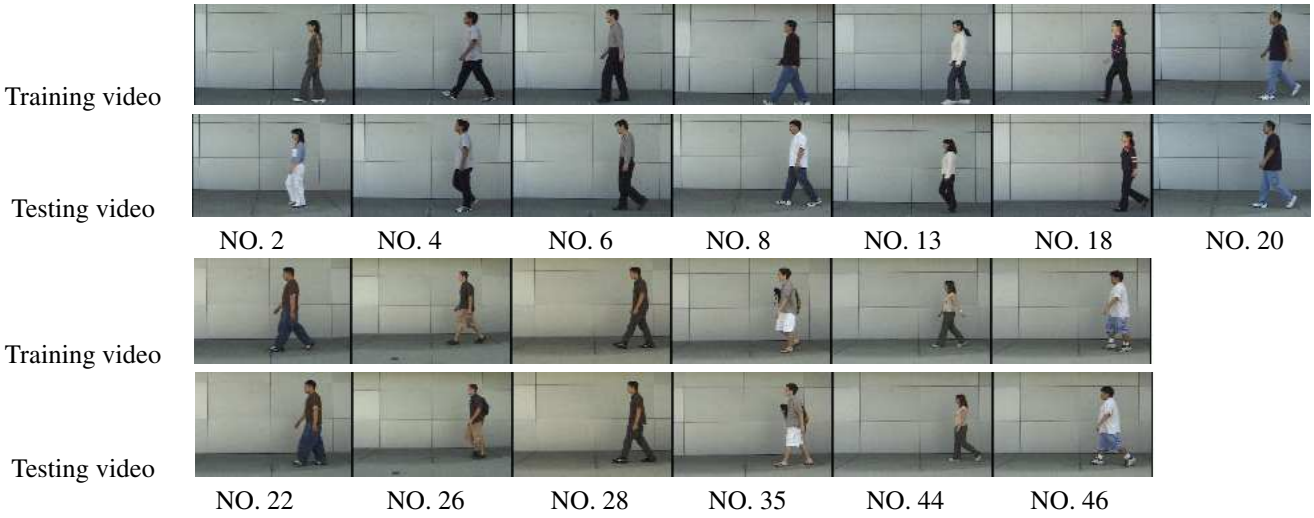


Figure 6. People misclassifi ed by OSFI only (see Table 2). For each person, one image of the training video sequence and one image of the testing video sequence are shown for comparison.

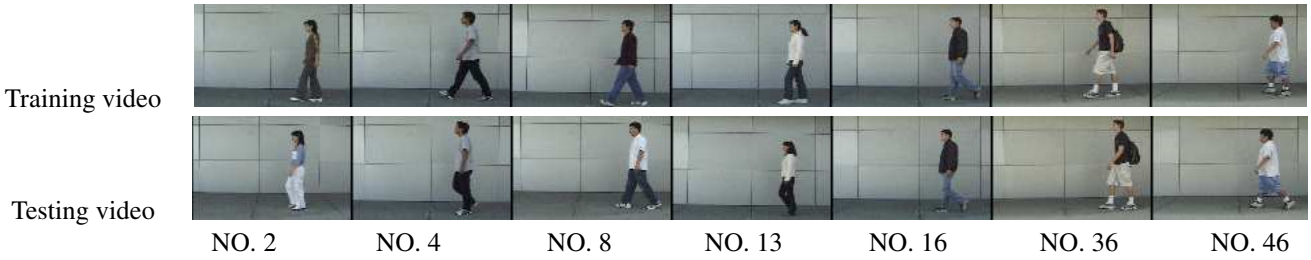


Figure 7. People misclassifi ed by ESFI only (see Table 2). For each person, one image of the training video sequence and one image of the testing video sequence are shown for comparison.

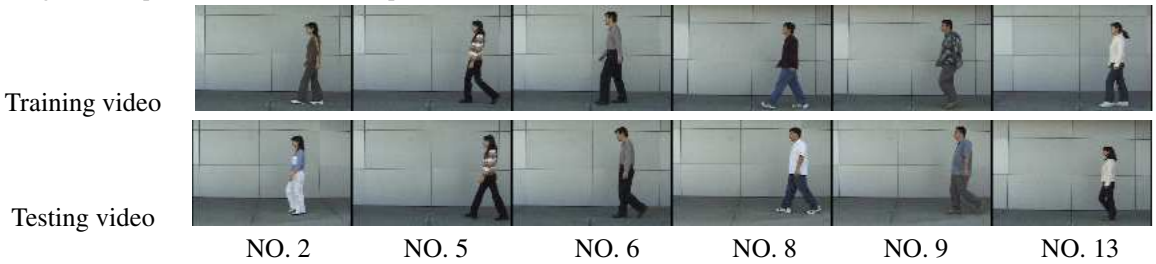


Figure 8. People misclassifi ed by GEI only (see Table 2). For each person, one image of the training video sequence and one image of the testing video sequence are shown for comparison.

Table 4. Q statistics

Fusion Method	N_11	N_00	N_01	N_10	Q Statistic
OSFI & GEI	31	4	9	2	0.75
ESFI & GEI	36	3	4	3	0.8

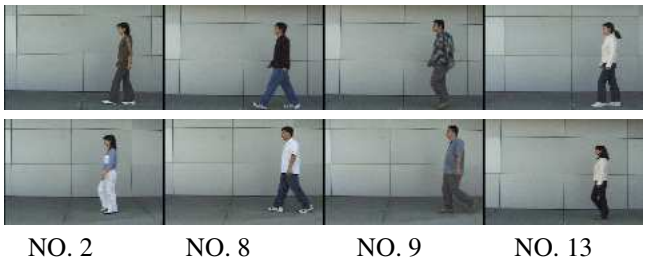


Figure 9. People misclassifi ed by ESFI and GEI using Sum rule (see Table 3). For each person, one image of the training video sequence and one image of the testing video sequence are shown for comparison.

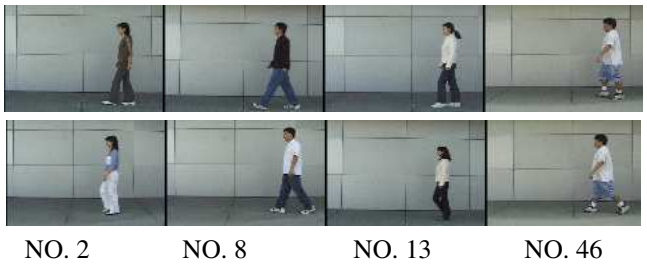


Figure 10. People misclassifi ed by ESFI and GEI using Max rule (see Table 3). For each person, one image of the training video sequence and one image of the testing video sequence are shown for comparison.

tion and classification are the same for ESFI and OSFI.

3.2. Performance Analysis

From Table 2, we can see that 71.8% people are correctly recognized by OSFI (13 errors out of 46 people), 84.8% people are correctly recognized by ESFI (7 errors out of 46 people) and 87.0% people are correctly recognized by GEI (6 errors out of 46 people). Among performance of fusion in Table 3, Sum rule and Max rule based on ESFI and GEI perform the best at the recognition rate 91.3% (4 errors out of 46 people). When Product rule is used, fusion does not improve the performance compared with the individual classifiers. For fusion based on OSFI and GEI, the best performance is achieved by Sum rule at 89.1% (5 errors out of 46 people), while for Product rule and Max rule, no improvement is obtained. Moreover, fusion based on ESFI and GEI always has better performance than fusion based on OSFI and GEI, except using Product rule where they are the same at 84.8% (7 errors out of 46 people). These results demonstrate the importance of constructing ESFI. From ESFI, we extract face features with more discriminating power. Therefore, better performance is achieved when ESFI instead of OSFI is used for recognition.

Performance of fusion can be analyzed more insightfully by the error index. Figure 6 shows people (video sequences) misclassified by OSFI only. Figure 7 shows people (video sequences) misclassified by ESFI only. Figure 8 shows people (video sequences) misclassified by GEI only. Figure 9 and Figure 10 show people (video sequences) misclassified by ESFI and GEI using Sum rule and Max rule, respectively. In Table 2, there are three misclassified people $\{2, 8, 13\}$ overlapped between classification using ESFI only and GEI only. There are four misclassified people $\{2, 6, 8, 13\}$ overlapped between classification using OSFI only and GEI only. From Table 3, we can see that the set of misclassified people $\{2, 8, 13\}$ are always a subset of the error indices when ESFI and GEI are combined by any fusion rule. Similarly, the set of misclassified people $\{2, 6, 8, 13\}$ are always a subset of the error indices when OSFI and GEI are combined by any fusion rule. These demonstrate that the match score fusion can not rectify the misclassification conducted by both of the face classifier and the gait classifier. People misclassified by the individual classifiers are likely to be classified correctly after fusion on the condition that there is at least one of the two classifiers works correctly.

For the performance improvement by fusion compared with the individual biometric, if the different classifiers misclassify features for the same person, we do not expect as much improvement as in the case where they complement each other [11]. We use a statistic to demonstrate that. There are several methods to assess the interrelationships between the classifiers in a classifier ensemble [15][3]. Given classifiers i and j corresponding to feature vectors f_i and f_j from the same person, respectively. We compute Q statistic:

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{9}$$

where N^{00} is the number of misclassification by both i and j; N^{11} is the number of correct classification by both i and j; N^{10} and N^{01} are the number of misclassification by i or j, but not by both. It can be easily verified that $-1 \leq Q \leq 1$. The Q value can be considered as a correlation measure between the classifier decision. The best combination is the

one that minimizes the value of Q statistic, which means the smaller the Q value is, the more potential the performance improvement by fusion has.

Table 4 shows the Q values using the classifiers based on OSFI and GEI, and the classifiers based on ESFI and GEI. The Q values using OSFI and GEI is 0.75, smaller than using ESFI and GEI at 0.8. The expected performance improvement using OSFI and GEI is higher than using ESFI and GEI. Experiments show that. For example, when Sum rule is used, the performance increase by fusion of OSFI and GEI is 17.4% (from 71.7% to 89.1%), while the performance increase by fusion of ESFI and GEI is only 6.5% (from 84.8% to 91.3%). Further, these two Q values are positive and relatively high, which indicate that many times the gait classifier and the face classifier are both performing correct classification or incorrect classification for the same person. In spite of this, we can see that our video based fusion system using ESFI and GEI is very promising since the fusion system, using Sum rule and Max rule, has better performance than either of the individual classifier. The best performance of fusion is 91.3%, i.e., 42 out of 46 people are correctly recognized using Sum rule and Max rule.

From the experiments, we can see the fusion system has better performance than either of the individual classifier when the appropriate fusion method is chosen. There are some people who are not correctly recognized by gait, but when side face information is integrated, the recognition rate is improved. It is because clothes or walking styles of these people are much different between the training and testing video sequences, so the gait classifier can not recognize them correctly. However, the side faces of these people don't change so much in the training and testing sequences. It shows that side face is a useful cue for the fusion system. On the other hand, since the face classifier is comparatively sensitive to the variation of facial expression and noise, it can not get a good recognition rate by itself. When gait information is combined, the better performance is achieved. Our experimental results demonstrate that the fusion system using side face and gait has potential since it integrates cues of side face and cues of gait reasonably, which are two complementary biometrics. Consequently, the fusion system is relatively robust compared with the system using only one individual biometric.

4. Conclusions

In this paper, an innovative video based fusion system is proposed, aiming at recognizing non-cooperating individuals at a distance in a single camera scenario. Information from two biometric sources, side face and gait, from the single camera video sequence, is combined using different fusion methods. The experimental results show that it is effective to integrate information from side face and gait for human recognition in video. ESFI constructed from multiple frames is a better face template than OSFI from a single frame. The best performance is 91.3%, archived by using Sum Rule and Max rule to integrate information from ESFI and GEI.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Recognizing people by their gait: the shape of motion. *Videre: Journal of Computer Vision Research*, 1(2):1–32, 1998. 4

[2] B. Bhanu and X. Zhou. Face recognition from face profi le using dynamic time warping. In *17th International Conference on Pattern Recogntion*, volume 4, pages 499–502, 2004. 3

[3] G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005. 7

[4] T. F. Chang. An improved algorithm for computing the singular value decompostion. *ACM Trans. Math. Softw*, 8:72–83, 1982. 4

[5] J. Han and B. Bhanu. Performance prediction for individual recognition by gait. *Pattern Reconition Letters*, 26, 2005. 3

[6] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Trans. PAMI*, 28(2):316–322, 2006. 3

[7] P. A. Hewitt and D. Dobberfuhl. The science and art of proportionality. *Science Scope*, pages 30–31, 2004. 3

[8] P. S. Huang, C. J. Harris, and M. S. Nixon. Recognizing humans by gait via parameteric canonical space. *Artificial Intelligence in Engineering*, 13:359–366, 1999. 4

[9] M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion and transparency. *Journal of Visual Communication and Image Representation*, 4:324–335, 1993. 3

[10] A. Kale, A. Roychowdhury, and R. Chellappa. Fusion of gait and face for human identifi cation. In *Proc. Acoustics, Speech, and Signal Processing 2004*, volume 5, pages 901–904, 2004. 1

[11] T. Kinnune, V. Hautamaki, and P. Franti. Fusion of spectral feature sets for accurate speaker identifi cation. In *Proc. 9th International Conference Speech and Computer (SPECOM'2004)*, pages 361–365, September 2004. 7

[12] J. J. Little and J. E. Boyd. Recognizing people by their gait: the shape of motion. *Videre: Journal of Computer Vision Research*, 1(2):1–32, 1998. 3

[13] G. Shakhnarovich and T. Darrell. On probabilistic combination of face and gait cues for identifi cation. In *Proc. Automatic Face and Gesture Recognition 2002*, volume 5, pages 169–174, 2002. 1

[14] G. Shakhnarovich, L. Lee, and T. Darrell. Integrated face and gait recognition from multiple views. In *Proc. Computer Vision and Pattern Recognition 2001*, volume 1, pages 439–446, 2001. 1

[15] C. A. Shipp and L. I. Kuncheva. Relationships between combination methods and measures of diversity in combining classifi ers. *Information Fusion*, 3, 2002. 7

[16] X. Zhou, B. Bhanu, and J. Han. Human recognition at a distance in video by integrating face profi le and gait. In *AVBPA*, pages 533–543, 2005. 1, 2